DOCUMENT RESUME

ED 320 954 TM 015 238

AUTHOR Shwalb, Barbara J.; Shwalb, David W.
TITLE The Design of a College Course Ratings Form by a Psychology "Test & Measurements" Class.
PUB DATE Apr 90
NOTE 21p.; Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Attitude Measures; College Faculty; *College Students; *Course Evaluation; Data Collection; *Evaluation Methods; Higher Education; Measurement; Psychology; Questionnaires; *Rating Scales; *Test Construction

ABSTRACT

Thirty psychology majors in a Tests and Measurement class, and 49 of 65 faculty members of a liberal arts arts college developed a college-wide course ratings questionnaire. The psychology students, in class exercises and homework assignments, collected data over a 6-month period from peers (n=153) and faculty at every stage of the design process. Free responses generated 654 rating items, and students sorted these into 22 theory-based categories. Several waves of systematic student and faculty screening reduced the item pool to 22 items. A 15-item questionnaire was finalized through factor analysis of student responses, and its validity and reliability were assessed before approval of the new form by the general faculty and administration. The Tests and Measurement class discussions centered on technical and conceptual issues pertinent to each phase of the research. The successful development and general acceptance of the new form were credited to student and faculty ownership in the design process. Two tables present study data. (Author/SLD)

The Design of a College Course Ratings Form by a Psychology "Test & Measurements" Class

Barbara J. Shwalb
University of Utah
David W. Shwalb
Westminster College of
Salt Lake City

Paper presented at the annual American Educational Research Association meetings; April 16-20, 1990; Boston.

Running Head: Tests & Measurements

Abstract

Thirty psychology majors in a Tests & Measurement class, and 49 of the 65 faculty of a liberal arts college developed a college-wide course ratings questionnaire. The psychology students, in class exercises and homework assignments, collected data from peers and faculty at every stage of the design process. Free responses generated 654 rating items, and students sorted these into 22 theory-based categories. Several waves of systematic student and faculty screening reduced the item pool to 22 items. A 15-item questionnaire was finalized through factor analysis of student responses, and its validity and reliability were assessed before approval of the new form by the general faculty and administration. The Tests & Measurements class discussions centered on technical and conceptual issues pertinent to each phase of the research. The successful development and general acceptance of the new form were credited to student and faculty ownership in the design process.

The Design of A College Course Ratings Form

by a Psychology "Tests & Measurements" Class

Student ratings of instructors have been researched more than any other form of faculty evaluation (Cohen, 1980) and yet this assessment method remains "controversial" and "widely criticized" (Marsh, 1984). One reason for this controversy is that many of the hundreds of student rating forms available have developed haphazardly and unscientifically. As a result, such questionnaires often have little face validity, so that instructors tend to view course evaluations with more suspicion than respect (Feldman, 1988). At the college at which this research took place, evaluation of teaching effectiveness had been based for over a decade on a six-item questionnaire.[1] Most faculty viewed the form skeptically as an administration tool used for promotion/salary decisions, and most students and faculty considered it inadequate.

The need to design a new ratings instrument provided an opportunity to enliven a course in "Tests & Measurements" (T&M) required of and unpopular among undergraduate psychology majors. In addition to involving the general faculty at every stage in the construction of this form, a class of thirty T&M students was enlisted to accomplish this task as a semester-long class exercise. The process by which students and faculty created the new questionnaire is detailed as follows.

Methods and Results

Overview of Data Collection

The following data were collected over a six month period, as a pool of 654 items was reduced to a finalized fifteen-item questionnaire:

1. Lists of free responses by faculty and students suggested items of importance for course/instructor ratings.
2. Following a systematic reduction of the pool to 66 items, faculty and students ranked 22 trios of statements. This enabled us to identify the single best item representing each of 22 theory-based categories.
3. Students rated several instructors on a six-point scale, using the 22 chosen items.
4. Faculty checked off what they considered the "five best" and "five worst" items from the same list of 22 items.

We next describe the instrumentation process in detail.

Free Responses

Thirty undergraduates in the T&M class were asked to write any five items they "would consider valuable for use in a new course evaluations form" during class time. For a homework assignment each student collected free responses from four other students (non-psychology majors). At the first general faculty meeting of the semester, 49 of the 65 faculty completed the same free-response questionnaire. In total, 49 faculty and 153 students (representing every academic division of the college) generated 654 items. T&M students printed each of these items onto a separate 3x5 index card, as their second assignment.

Theory-Based Categorization of Items

The nineteen dimensions of course ratings extracted from the research literature by Feldman (1987) served as categories sort the free-response items. Informal screening of the 654 items by the two authors, and suggestions by faculty colleagues, indicated the need for three additional categories (#20 "Difficulty Level of the Course", #21 "General Recommendation of Course/Instructor" and #22 "Text and Assignments"). One of the authors sorted the cards into these 22 categories and the T&M students performed the same procedure at a class meeting, as a measure of inter-rater reliability. The instructor and students sorted 508 (78%) of the items into the same categories with a range of concordance from 33% (for "Value of Course beyond Academics") to 100% (for "Intellectual Challenge"). The category name with Feldman's (1987) dimension number in parentheses, the number of index cards sorted into each category by the instructor, and the inter-rater concordance rate is given in the last three columns in Table 1. Sixty-four items were not codable into any of the 22 categories.

Insert Table 1 about here

Screening of Items by Students and Faculty

Next, 337 similarly-worded item cards were screened out. The remaining 253 items were typed onto a 22-page questionnaire, listing on each page the items for each category. This form was submitted for screening to an eight-member faculty committee which had been created the previous year to devise a new ratings questionnaire. This panel reached a consensus on the "three most significant" items representing each of the 22 dimensions.

T&M students then completed a questionnaire in class ranking each of 22 trios of items from "1" to "3" in order of significance, and distributed this form to 100 other undergraduates as their third assignment. Forty-one faculty members completed and returned the same form through campus mail. Rankings were used to determine the single best item under each dimension, and the highest-ranked of faculty and students coincided for 18 of the 22 categories, a higher degree of student/faculty concordance than is often found in the research literature (Feldman, 1988). In the cases of the other four dimensions (Feedback, Fairness, Personality and Organization), the faculty's first choices were used for the next round of testing.

Pilot-Testing and Factor Analysis

The 22 items selected were next pilot-tested on students and assessed further by the faculty. In the final month of the semester students in the T&M and four other classes rated their courses twice on the same 22 items, using a 6-item scale (1 = strongly disagree; 6 = strongly agree), with a ten-day interval between administrations. The test-retest correlation of ratings by these 103 students ranged from .31 (Workload) to .78 (Interest Stimulation), with an average correlation of .58.

Data from first administration of the test-retest ratings, and those by 172 students from eleven other classes representing all divisions of the college, were factor-analyzed using the Varimax procedure. Factor analysis described the overall dimensions by students, and was used to finalize the questionnaire. Factor loadings and test-retest reliability are reported in the first six columns of Table 1. Items are grouped in Table 1 according to the factors extracted, and those items which were subsequently eliminated from the final version of the questionnaire are listed last as "Eliminated Items".

Analyses extracted four factors meeting the criterion Eigen value of 1.00: The first factor, "Dynamic/Communication" (six items), "Organization" (three items) "Concern for Individuals" (three items), and "Outcomes" (two items). These four dimensions extracted here clearly replicated those found in several published studies of course rating "dimensionality" (Abrami, 1985; Beatty, Frey & Leonard, 1975; Hildebrand, Wilson & Dienst, 1971; Marsh, 1983; Warrington, 1973).

Faculty Finalization of the Questionnaire

At the next general faculty meeting, forty faculty members checked off ten of the same 22 items as either among the five "best" and five "worst" items, indicated how many items they wanted on the rating form, and voted for or against an open-ended question "additional comments" space on the form. Fifteen items reached the dual criteria for final inclusion on the ratings questionnaire: (1) a factor loading of .50 and (2) being rated among the "5 best" by more faculty than among the "5 worst". The mean desired number of items was 15.1 (mode = 15) and 80% of the faculty approved of space for "additional comments."

The wording and format of the new course ratings form was finalized by the eight-member faculty committee, and after minor revisions in the wording of some items, the general faculty unanimously approved the fall 1989 implementation of the new instrument. In total, development of the new college-wide ratings questionnaire required six months, at a total expense of $70 for computing costs.

Instrumentation as a Class Exercise

In addition to the above described service to the college, a major purpose of the research reported was to illustrate key concepts of instrumentation and statistics to the T & M class, through class activities and homework assignments. Class lectures and discussion related each activity to major testing/measurement concepts contained in the course's textbook.

Table 2 lists the major topics discussed in class, in relation to each stage of the research. One example of a class activity is as follows. After T & M students listed their free-response items, the instructor led a discussion of how ecological validity is enhanced when subjects from the target population generate the questionnaire items. The students critiqued the free-response format from the viewpoint of the respondent, and were told to keep these criticisms in mind as they sampled friends' free responses as a homework assignment. The following week, they discussed the difficulties involved in locating cooperative respondents among their friends, and in gaining serious compliance from respondents. The general issues of content/ecological validity, sampling and test administration were included in the text readings and class lecture that week, and a 16-mm film on test administration further reinforced these materials.

Insert Table 2 about here

Discussion

T & M Students' Responses to Their Experience

The 30 T & M psychology majors participated in the study as researchers and subjects, at every stage of the research. This was the first practical research experience for most of the students, and reactions to this activity indicated that it enhanced their overall satisfaction with the study of tests and measurements.

At the conclusion of the course, students answered a questionnaire about the instructional value of the text design experience. On a six-point scale (1 = didn't add anything; 6 = really added to my understanding) students rated the value of the experience for their understanding validity, reliability, factor analysis, test construction, population sampling, and test administration. The mean rating for the six items was 4.68, with 86% of the respondents rating all items on the positive half of the scale (4, 5, or 6). Clearly, the students saw their participation in the research as a productive learning experience. Because the small psychology department normally provides minimal research experiences (more commonplace in a university department), this hands-on opportunity was especially valuable. The involvement of students in the instrumentation process also contributed to general acceptance of the new questionnaire by the college studentbody. Informal questioning of several students after the initial college-wide implementation of the ratings form in the fall of 1989, indicating just such a favorable response to the new items and response format.

Faculty Involvement

A year before the initiation of this research, the general faculty had formally requested that a new ratings form be devised, but little progress had been made towards that end. An ad-hoc faculty committee produced a twenty-item form after a few brainstorming and working sessions, but this form was discarded by the general faculty as unacceptable. The major criticism of this initial aborted effort was that items were "just thrown together, with no rhyme or reason," in the words of one faculty member. One aim of the present study was to produce a questionnaire acceptable to the faculty, by bringing more faculty into the design process and utilizing systematic procedures.

The new form was developed in close collaboration with a majority of the college's teaching faculty and faculty input was frequently solicited. As noted by Kulik (1977, p. 2), "one problem with the traditional approach to course evaluations is that teachers do not become highly involved in the rating process when they collect data on forms created by someone else for someone else's purposes." As a result of this distance between instructors and the ratings-form process, many faculty are negative towards course evaluations, and course evaluations are misunderstood despite over 1300 research studies on the subject (Cashin, 1988). It is possible that the general faculty's approval of the new form developed here was partially a result of their extensive involvement in the design process. And though faculty ownership of the form was emphasized, most faculty were asked in total for less than an hour of their time. In addition, the use of computer analyses and established social scientific methods impressed the faculty further with the legitimacy of the project. This study showed that by involving faculty and students, scientific procedures can be successfully

implemented at a negligible cost. In the present study, we improved a system of college-wide instructor ratings, and made a "Tests and Measurements" course come alive.

References

Abrami, P. C. (1985). Dimensions of effective college instruction. Review of Higher Education, 8, 211-28.

Cashin, W. E. (1988). Student ratings of teaching: A summary of the research. Kansas State University: Center for Faculty Evaluation and Development, Idea Paper No. 20.

Cohen, P. A. (1980). Effectiveness of student-rating feedback for improving college instruction: A meta-analysis of findings. Research in Higher Education, 13, 321-341.

Feldman, K. A. (1987). Research productivity and scholarly accomplishment of college teachers as related to their instructional effectiveness: A review and exploration. Research in Higher Education, 26, 227-298.

Frey, P. W., Leonard, D. W., & Beatty, W. W. (1975). Student ratings of instruction: Validation research. American Education Research Journal, 12, 327-336.

Feldman, K. A. (1988). Effective college teaching from the students' and faculty's view: Matched or mismatched priorities. Research in Higher Education, 28,(4), 291-344.

Frey, P. W., Leonard, D. W., & Beatty, W. W. (1975). Student ratings of instruction: Validation research. American Educational Research Journal, 12, 327-336.

Hildebrand, M., Wilson, R. C., & Dienst, E. R. (1971). Evaluating university teaching. Berkeley: Center for Research and Development in Higher Education, University of California, Berkeley.

Kulik, J. A. (1977). Adapting flexible instructional evaluation to another faculty, another campus. Symposium paper presented at the 1977 Annual Meetings of the American Educational Research Association, New York.

Marsh, H. W. (1983). Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics. Journal of Educational Psychology, 75, 150-166.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. Journal of Educational Psychology, 76(5), 707-754.

Warrington, W. G. (1973). Student evaluation of instruction at Michigan State University. In A. L. Sockoff (Ed.), Proceedings: The first invitational conference on faculty effectiveness as evaluated by students (pp. 164-182). Philadelphia: Temple University, Measurement and Research Center.

Footnote

[1]The six items in the previously used questionnaire were:

1. What is your overall rating of the instructor?
2. Does the instructor communicate interest and enthusiasm in the subject?
3. Does the instructor inspire confidence in her/his knowledge of the subject matter?
4. Has the instructor been genuinely concerned with students' progress and been actively helpful?
5. Does the instructor use effective methods for presenting material and concepts?
6. Would you recommend this course to your fellow students?

The items were rated on a 7 point scale (1 = "low", 7 = "high").

Author Notes

This paper is based on a presentation at a Division J session at the Boston 1990 annual meeting of the American Educational Research Association. The thoughtful participation of the students, faculty, and administrators of Westminster College of Salt Lake City is gratefully acknowledged. Dr. Ralph Reynolds and Dr. Steve Baar provided additional support. Address reprint requests to: B. Shwalb, University of Utah, Educational Studies, 115 MBH, Salt Lake City, UT 84112 or D. Shwalb, Psychology Department, Westminster College, Salt Lake City, UT 84105.

Table 1
Factor Loadings, Reliability Coefficients, Faculty Evaluations and Underlying Constructs of 22 Rating Items

	Factor Loadings 1	2	3	4	Test-Retest Reliability	% Faculty Approval	Underlying Construct Feldman (1987) Dimension	# of Items Sorted	Intersorter Reliability
Dynamic/Communication Items (6)									
The course stimulated my interest in the subject.	.81	.15	.25	.26	.78	.67	Interest Stimulation (1)	24	.96
The instructor communicated enthusiasm about the subject.	.81	.20	.25	.12	.77	.55	Enthusiasm (2)	31	.81
The instructor presented material clearly.	.75	.24	.20	.24	.58	.64	Presentation Clearness (6)	52	.90
The instructor encouraged useful questions and discussion.	.65	.24	.27	.13	.54	.86	Encouraged Discussion (16)	24	.71
The course challenged me to think.	.69	.14	.29	.33	.55	.83	Intellectual Challenge (17)	18	1.00
I received useful feedback about my performance.	.50	.27	.54	.07	.52	.63	Feedback (15)	8	.50
Organization Items (3)									
The instructor was prepared when presenting material.	.53	.69	.11	.08	.50	.79	Preparedness (5)	23	.74
The instructor was knowledgeable about the subject	.63	.61	.03	.11	.41	.82	Subject Knowledge (3)	39	.85
I understood the course objectives and the instructor's expectations.	.29	.67	.33	.19	.44	.90	Clarity of Objectives (8)	33	.76
Concern for Individuals Items (3)									
The instructor graded my performance fairly.	.13	.28	.50	.33	.67	.83	Fairness (13)	40	.85
The instructor showed interest in having students learn.	.50	.13	.64	.12	.69	.84	Sensitivity (7)	20	.70
The instructor was available during office hours or by appointment.	.06	.25	.75	.13	.54	.55	Helpfulness/ Availability (19)	23	.87

Table 1
(Continued)

	Factor Loadings 1	2	3	4	Test-Retest Reliability	% Faculty Approval	Underlying Construct Feldman (1987) Dimension	#of Items Sorted	Intersorter Reliability
Outcome Items (2)									
I increased my understanding of the subject.	.50	.22	.05	.55	.57	.59	Academic Value (19)	30	.63
The class was a valuable learning experience.	.67	.24	.12	.51	.59	.50	Overall Liking (new)	54	.96
Eliminated Items (7)									
The Instructor conducted him/herself professionally.	.56	.47	.24	.04	.64	.23	Personality (14)	14	.96
The course content was pertinent to my educational goals.	.12	.09	.16	.83	.62	.16	Impact (12)	24	.29
The instructor accomplished what he/she set out to do.	.54	.53	.28	.12	.71	.03	Organization (4)	17	.53
The instructor used simple, effective teaching methods.	.64	.31	.29	.14	.51	.30	Aids/ Materials (11)	18	.72
The workload was appropriate for the semester hours of the course.	.01	.61	.29	.32	.31	.32	Difficulty (new)	24	.92
The instructor facilitated linkages between readings and lectures.	.52	.50	.17	.25	.69	.23	Assignments (new)	46	.76
The course gave insight and knowledge for future use.	.48	.35	.21	.56	.59	.46	Value Beyond Academics (10)	21	.53
The instructor was concerned about student progress.[1]	.43	.04	.80	.06	.54	.70	Respect/ Concern (18)	13	.92

Note [1]This item was eliminated by the faculty panel and in discussion by the general faculty in the finalization of the form, and replaced by a new item, "Exams and assignments reinforced my learning," better reflecting faculty sentiment.

Table 2

Stages of the Instrumentation Process,

and Corresponding T & M Class Discussion Topics

Research Stage	Discussion Topics
Faculty/student free responses	Sampling a population Administering a test Content/ecological validity
Instructor choice of 22 categories	Construct validity Theory-based test construction
Student sorting 654 items into 22 categories	Inter-rater reliability Sorting methodologies
Faculty panel selection of 66 items	Expert panel ratings Objective vs. subjective measures
Student/faculty ranking of 66 items to reduce to 22 items	Rankings vs. ratings Mail vs. face-to-face surveys
Faculty ratings of 22 items	Self-report paper & pencil methodology Format of a questionnaire
275 students course ratings on 6-point scales	Test-retest reliability & correlation Factor analysis Likert scales and anchor points